Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **OCT 2014**	2. REPORT TYPE	3. DATES COVERED **00-00-2014 to 00-00-2014**

4. TITLE AND SUBTITLE **VOCALinc**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Massachusetts Institute of Technology,Lincoln Laboratory,244 Wood Street,Lexington,MA,02420**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S)
	11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

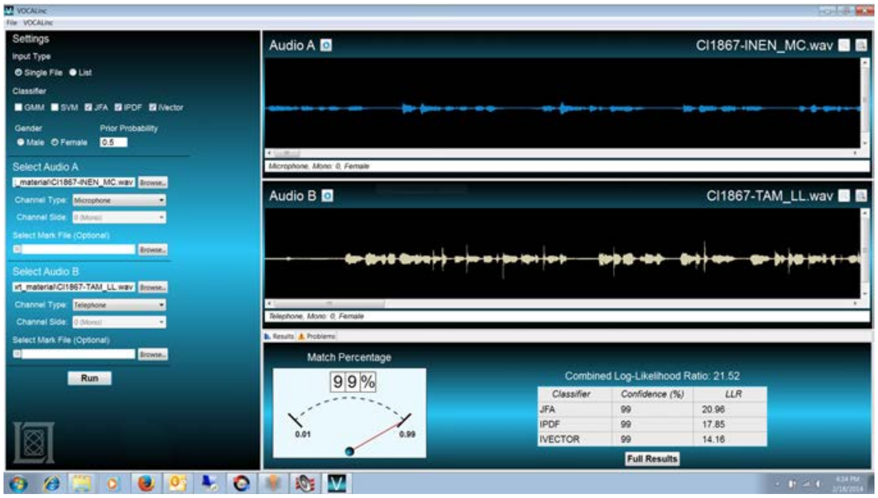
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**2**	

Tech Notes

www.ll.mit.edu October 2014

VOCALinc

Automated speaker recognition software provides an accurate, objective, consistent, and efficient tool for conducting speaker comparisons.



Settings
Input Type
○ Single File ● List
Classifier
■ GMM ■ SVM ☑ JFA ☑ IPDF ☑ IVector
Gender ● Male ○ Female
Prior Probability 0.5

Select Audio A
_material\CI1867-INEN_MC.wav
Channel Type: Microphone
Channel Side: 0 (Mono)
Select Mark File (Optional)

Select Audio B
rt_material\CI1867-TAM_LL.wav
Channel Type: Telephone
Channel Side: 0 (Mono)
Select Mark File (Optional)

Microphone, Mono: 0, Female

Telephone, Mono: 0, Female

Combined Log-Likelihood Ratio: 21.52

Classifier	Confidence (%)	LLR
JFA	99	20.96
IPDF	99	17.85
IVECTOR	99	14.16

VOCALinc's graphical user interface features checkboxes, buttons, and drop-down menus for analysts to input speaker gender, channel type, prior probability, and other metadata regarding the uploaded voice samples (labeled Audio A and Audio B above), as well as to choose the desired algorithm(s) for comparing and scoring the samples.

As part of a growing trend in biometric identification, speaker recognition is becoming increasingly important to the nation's forensic laboratories, intelligence agencies, and military branches. Voice, along with other biometrics, such as fingerprints and iris patterns, is a unique characteristic that can be used to verify or identify individuals. Until recently, only spectrogram-based speaker recognition techniques were available. These techniques rely on electronically recorded speech signals, graphically represented as a function of time and frequency (i.e., a spectrogram), that analysts must manually evaluate—an error-prone, subjective, inconsistent, and time-consuming process.

VOCALinc, a stand-alone automated speaker recognition software developed by MIT Lincoln Laboratory, overcomes many of the issues inherent in manual techniques. Incorporating recent advances in speaker recognition technology, the software consists of a suite of speaker recognition algorithms that conduct the speaker comparisons and a graphical user interface that allows analysts to specify known data about uploaded voice samples and apply the algorithm(s) of their choice. VOCALinc outputs two pieces of vital information:

- A match score between the voice samples in question provides an automatic assessment of the confidence level that the voices under evaluation were produced by the same speaker.
- A comprehensive report on the voice samples' signals (e.g., speech duration, signal-to-noise ratio, and speech clipping percentage) informs analysts of signal conditions that could potentially discount the results (e.g., short voice samples).

Combined, this information equips analysts with sufficient data to make an informed decision about speaker identity.

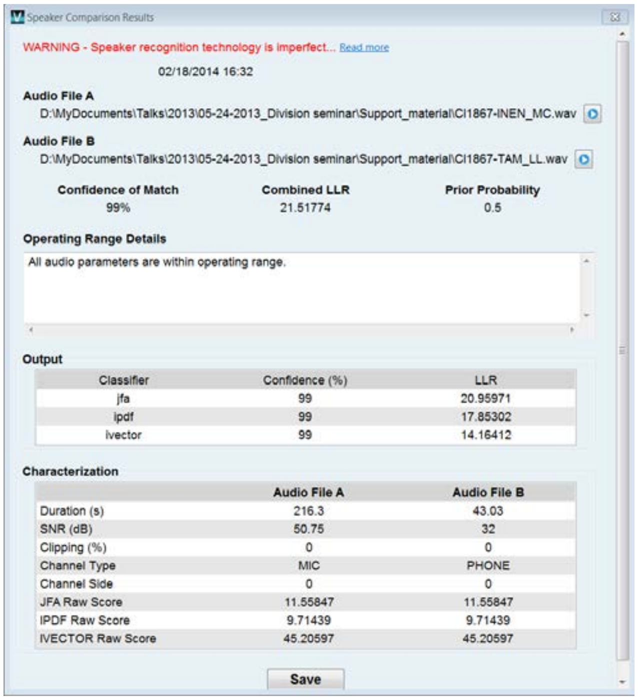
A Suite of Algorithms

VOCALinc computes speaker comparisons through a suite of automated speaker recognition classifiers, or algorithms, which can be divided into two main classes: legacy algorithms and state-of-the-art algorithms. Two legacy algorithms—Gaussian mixture model and support vector machine—enable analysts to assess results from VOCALinc in the context of results obtained from older speaker recognition techniques. These classifiers work well when the voice samples under analysis are in milder conditions, such as in low-noise environments and on telephone channels, but do not work well in cross-channel (e.g., microphone versus telephone) and cross-language (e.g., English versus Spanish) conditions, both of which are common in forensic and investigative cases. To address these channel variability issues, VOCALinc incorporates a newly developed class of interoperable techniques: inner product discriminant functions, joint factor analysis, and i-vector algorithms. Each algorithm is trained to reduce or eliminate sources of signal variation while preserving desirable speaker-specific characteristics.

After the algorithm(s) compute the comparisons, a final stage, known as backend or calibration, normalizes the output score to a 0–99% score. If an analyst selects more than one algorithm, each algorithm is scored independently and the scores produced are normalized. The backend, trained by using machine learning, also automatically recognizes score patterns across multiple algorithms and modifies the combined score. VOCALinc generates a final match score by fusing the normalized scores from each of the chosen clasifiers. This fusion uses a held-out data set (data that has not been used by the system for any other training purposes) along with the statistical techniques of hypothesis testing and Bayes' theorem to produce the final match score.

Capabilities of VOCALinc

VOCALinc supports language- and text-independent recognition, i.e., speakers can be identified regardless of language and speech content. In terms of processing capacity, the speaker recognition engines support both one-to-one and simultaneous comparisons on large datasets. More than 1,000,000 comparisons can be conducted in well under one hour, enabling analysts to quickly zero in on data of interest. For 60 seconds of speech under realistic conditions, VOCALinc can obtain a 95% or better match; under favorable conditions (i.e., without noise, channel differences, etc.), it can achieve 98% match confidence. VOCALinc can function in cross-language and cross-channel conditions with as little as 10 seconds of speech. Legacy and state-of-the-art speaker recognition algorithms were deliberately included in the suite to enable cross-checking of past results and to overcome channel and language variability issues. Applying more than one of these algorithms to a comparison improves speaker recognition accuracy by 5–20%, depending on the number and type of algorithms selected.



WARNING - Speaker recognition technology is imperfect... Read more

02/18/2014 16:32

Audio File A
D:\MyDocuments\Talks\2013\05-24-2013_Division seminar\Support_material\CI1867-INEN_MC.wav

Audio File B
D:\MyDocuments\Talks\2013\05-24-2013_Division seminar\Support_material\CI1867-TAM_LL.wav

Confidence of Match	Combined LLR	Prior Probability
99%	21.51774	0.5

Operating Range Details

All audio parameters are within operating range.

Output

Classifier	Confidence (%)	LLR
jfa	99	20.95971
ipdf	99	17.85302
ivector	99	14.16412

Characterization

	Audio File A	Audio File B
Duration (s)	216.3	43.03
SNR (dB)	50.75	32
Clipping (%)	0	0
Channel Type	MIC	PHONE
Channel Side	0	0
JFA Raw Score	11.55847	11.55847
IPDF Raw Score	9.71439	9.71439
IVECTOR Raw Score	45.20597	45.20597

A pop-up results window features a match percentage output score (ranging from 0–99%) for each speaker comparison algorithm selected and a detailed signal condition report to characterize each audio file. In this case, a one-to-one (1:1) comparison is conducted (between Audio File A and Audio File B), but VOCALinc is also capable of performing simultaneous speaker comparisons on massive datasets on the scale of 1000s of comparisons. This capability is important for cases in which an analyst may want to conduct several comparisons on multiple audio files before focusing on a smaller dataset.

Benefits of VOCALinc

VOCALinc was developed utilizing U.S. government operational data and with direct access to end-user feedback. As a result, its algorithms and interface were iteratively designed to ensure a best-practice approach. VOCALinc benefits its users by

- Reproducing consistent results among different laboratories
- Easing analyst workload and casework backlog
- Expediting open-case resolution
- Minimizing the number of false leads and the costs associated with pursuing them
- Being customizable to classified conditions
- Including an automated installation validator to calibrate speaker comparison equipment for quality-assurance purposes
- Supporting the ANSI/NIST-ITL Type-11 Record Standard[1]

With applications in intelligence missions concerning national security (e.g., terrorism), forensic investigations within the field of law enforcement (e.g., drug trafficking), and military deployments (e.g., force protection), VOCALinc has the potential to become a valuable tool in a variety of settings. In fact, the software is already in use by several entities. Future versions of VOCALinc could reduce memory usage and enhance robustness to unseen devices such as body microphones and multirecording systems. ■

Technical Points of Contact
Dr. Joseph P. Campbell
Human Language Technology Group
jpc@ll.mit.edu
781-981-7522

Dr. Pedro A. Torres-Carrasquillo
Human Language Technology Group
ptorres@ll.mit.edu
781-981-5581

For further information, contact
Communications and
Community Outreach Office
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420-9108
781-981-4204

[1] The ANSI/NIST-ITL (American National Standards Institute/National Institute of Standards and Technology-Information Technology Laboratory) Type-11 Record Standard defines the standard for the transmission of voice data and related information in forensic and invesigative speaker comparison.

This work is sponsored by the Department of Defense under Air Force Contract #FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the author and are not necessarily endorsed by the U.S. government.